Regression Modelling with Many Correlated Predictors:

A new approach to linear and logistic regression with high dimensional data

Jay Magidson and Gary Bennett

4th September 2012 – RSS Conference, Telford, UK

Statistical Innovations
Thinking outside the brackets!™

logit research

Tel: 01732 465662 e-mail: garyb@logitresearch.com

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Near Collinearity

- The matrix formula for **b** is give by: $\mathbf{b} = \mathbf{(X'X)^{-1}X'Y}$
- **X** = Predictor matrix of dimension **N x P**
- Assumes no exact linear relationships among the P predictors in the model
  - Otherwise the Matrix **X'X** is singular and OLS Regression cannot occur
- With **near**-collinearity of **X** (near singularity of **X'X**), get imprecise, unstable parameter estimates
  - When tolerance or N-P are small, small data changes produce wide swings in the parameter estimates and large instability with respect to prediction on new cases.
  - Coefficients may have very high standard errors and still be statistically significant because of implausible (high) magnitudes.
  - Coefficients might have the "wrong" sign or signs that change as another variable enters or is removed from model
  - As P approaches N data becomes increasingly collinear and R^2 tends to 100% - EVEN WHEN POPULATION PREDICTORS ARE INDEPENDENT!

Many "dabblers" in stats fooled into thinking they are getting good models!

Statistical Innovations Thinking outside the brackets!™ logit research

Standard Errors on Parameters

- The Standard Error of the parameter (b_g) of the gth coefficient is given by the formula

$$\mathbf{Var}(b_g) = \frac{1}{N-P} * \frac{s_Y{}^2}{s_g{}^2} * \frac{(1-R^2)}{(1-R_g^2)}$$

s_{Y^2}=variance of Y, s_{g^2}=variance of x_g

$1 - R_g^2$ = the tolerance of x_g where R_g^2 is from the regression of x_g on the other predictors

- Note:
  - Variance of the bs is high for small N
  - As P (no. of predictors) approaches N variance among the bs tends to infinity (formula assumes N≥P)
  - Higher correlation among the Xs leads to lower denominator and higher variance among the bs

Statistical Innovations Thinking outside the brackets!™ logit research

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Solution: Impose Regularisation

- Set one or more regression coefficients to zero:
  - eliminating extraneous predictors maintains unbiasedness and reduces variance, thus reducing prediction error.
- Penalised regression – restrict magnitude of regression coefficients, biasing them towards zero, but reducing variance
  - Ridge Regression, Lasso, Elastic Net (GLMNET).
- Component/Dimension reduction strategies – set effects of higher dimensions to zero, thus reducing variance.
  - In practice this is done by replacing the P predictors x_g, $g=1,2,\ldots,P$ by $K \leq P$ components:
    - Principal Components Regression (PCR)
    - Partial Least Squares Regression (PLS-R)
    - Correlated Component Regression (CCR)

Statistical Innovations Thinking outside the brackets!™ logit research

CCR – How are Components Formed?

- CCR models are based on two tuning parameters:
  - K = Number of components, P = Number of Predictors*
- As with PCR and PLS-R each component K is an exact linear combination of the included predictors
  - The weights in CCR are chosen to maximise the components ability to predict y
  - Each of the K Components can be thought of as a composite predictor
  - Regression weights are obtained for the components, which can be used to obtain regression coefficients for the predictors.
- EXAMPLE K=2, P=10:
  - A standard regression (no components) would yield an intercept + 10 coefficients
  - CCR with K=2 yields an intercept + 2 **component weights**
  - Since components can be expressed in terms of predictors, a reduced form results in an intercept +10 **regularized coefficients** for predictors.

* All P predictors included in model unless Stepping Down procedure is enabled

Statistical Innovations Thinking outside the brackets!™ logit research

CCR – Regularized Regression Coefficients

1st Component: $S_1 = \sum_{g=1}^{P} \lambda_g^{(1)} x_g = \lambda_1^{(1)} x_1 + \lambda_2^{(1)} x_2 + \ldots + \lambda_P^{(1)} x_P$

2nd Component: $S_2 = \sum_{g=1}^{P} \lambda_g^{(2)} x_g = \lambda_1^{(2)} x_1 + \lambda_2^{(2)} x_2 + \ldots + \lambda_P^{(2)} x_P$

Regression K=2: $\hat{y} = \alpha_0 + b_1 S_1 + b_2 S_2$ b_1 and b_2 are **component weights**

Regression Decomposed:

$$\hat{y} = \alpha_0 + b_1\lambda_1^{(1)} x_1 + b_1\lambda_2^{(1)} x_2 + \ldots + b_1\lambda_P^{(1)} x_P + b_2\lambda_1^{(2)} x_1 + b_2\lambda_2^{(2)} x_2 + \ldots + b_2\lambda_P^{(2)} x_P$$

$\boldsymbol{\beta_g}$ **is regularized coefficient**

$$= \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_P x_P \quad \text{where} \quad \boldsymbol{\beta_g = b_1\lambda_g^{(1)} + b_2\lambda_g^{(2)}}$$

This applies for any value of K or P (K≤P)
NOTE K=P (SATURATED MODEL) GIVES SAME COEFFICIENTS AS A STANDARD OLS REGRESSION

Statistical Innovations Thinking outside the brackets!™ logit research

Algorithm: Computation of S_1 – Component 1

- The first component (S_1) is simply a weighted average of all P one predictor effects from regressions of y on x.
- Simply estimate P one predictor models (simple regressions):

$$\hat{y} = \alpha_g^{(1)} + \lambda_g^{(1)} x_g \quad g = 1,2,\ldots,P$$

- We then form the first component S_1 as a linear combination of X using the parameters from the 1-predictor models as weights):

$$S_1 = (1/P) \sum_{g=1}^{P} \lambda_g^{(1)} x_g \qquad \hat{Y}^{(1)} = a + b_1 S_1$$ 1-component model

This component captures the direct effects of X for the regression y on X as a simple average of the one predictor models

Algorithm: Computation of S_2 – Component 2

- If K>1, S_2 is defined as a weighted average of P *partial* effects:

$$\lambda_g^{(2)} x_g \quad g = 1,2,\ldots,P$$

- Where each $\lambda_g^{(2)}$ is estimated from the following 2-predictor model:

Covariate

$$\hat{y} = \alpha_g^{(2)} + \gamma_{1.g}^{(2)} S_1 + \lambda_g^{(2)} x_g \quad g = 1,2,\ldots,P$$

- We then form the second component S_2 as a linear combination of X using the parameters from each of the models as weights:

$$S_2 = (1/P) \sum_{g=1}^{P} \lambda_g^{(2)} x_g$$

$$\hat{Y}^{(2)} = a + b_1 S_1 + b_2 S_2$$ 2-component model

This is essentially the same process as used to derive the loading $\lambda_g^{(1)}$ for predictor g, except that the contribution of S_1 is partialled out by using it as a **covariate** in each simple regression to calculate the loading $\lambda_g^{(2)}$

Statistical Innovations Thinking outside the brackets!™ logit research

Algorithm: Weights for S_3 – Component 3

- If K>2, S_3 is defined as a weighted average of the P *partial* effects:

$$\lambda_g^{(3)} x_g \quad g = 1,2, \dots, P$$

- Where each $\lambda_g^{(3)}$ is estimated from the following 3-predictor model:

Covariates

$$\hat{y} = \alpha_g^{(3)} + \gamma_{1.g}^{(3)} S_1 + \gamma_{2.g}^{(3)} S_2 + \lambda_g^{(3)} x_g \quad g = 1,2, \dots, P$$

- We then form the third component S_3 as a linear combination of X using the parameters from each of the models as weights:

$$S_3 = \sum_{g=1}^{P} \lambda_g^{(3)} x_g \qquad \hat{Y}^{(3)} = a + b_1 S_1 + b_2 S_2 + b_3 S_3$$

3-component model

This is essentially the same process as used to derive the loading $\lambda_g^{(2)}$ for predictor g, except that the contribution of **both** components (S_1 and S_2) are partialled out in each single regression to calculate the predictor loading $\lambda_g^{(3)}$.

Statistical Innovations Thinking outside the brackets!™ logit research

Interpretation of Components

- S_1 captures direct effects of predictors on y
- S_2, correlated with S_1, captures the effects of suppressor variables that improve prediction by removing extraneous variation from one or more of the predictors that have direct effects
- S_3 and higher components can also be interpreted as capturing suppressor effects
- Can divide predictors into two types:
  - Prime predictors (those having direct effects) are identified as those having substantial loadings on S_1
  - Proxy predictors (suppressor variables) as those having substantial loadings on S_2+, and relatively small loadings on S_1
- It is possible to rotate the components to improve interpretation if required

Statistical Innovations Thinking outside the brackets!™ logit research

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Optimising K and P – Optimal Predictors unknown

- Works in conjunction with **M-fold cross-validation (CV)** to select the value P(K) = predictors at a **user-specified value of K**
- Step 1:
  - Initially use CV to estimate K-component CCR model based on all P predictors.
  - Different P-predictor submodels estimated for each fold eliminated and each used to score omitted fold. **CV performance (R^2) for P predictors** is then obtained based on predictions from all eliminated folds combined.
- Step 2:
  - For each submodel coefficients are standardised and the predictor with the **smallest absolute value** of its standardised coefficient (i.e., the least important predictor) is eliminated.
  - Different P-1 predictor submodels estimated for each fold eliminated and each used to score omitted fold. **CV performance (R^2) for P-1 predictors** is then obtained based on predictions from all eliminated folds combined.
- Return to Step 2 reducing the number of predictors by 1 until P=1
- Choose P(K) as value for P that yields the highest CV**(R^2)**
- Estimate **K-component model** on entire (training) sample; step down to P(K) predictors.

Statistical Innovations Thinking outside the brackets!™ logit research

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Example 1: Predicting Car Price

N=24 car models; P=6 predictors
dependent variable: PRICE = price of a car (measured in francs)

Each predictor is positively correlated with PRICE.

Explanatory Variable	Correlation with PRICE
CYLINDER (engine measured in cubic centimeters)	.85
POWER (horsepower)	.89
SPEED (top speed in kilometers/hour)	.72
WEIGHT (kilograms)	.81
LENGTH (centimeters)	.75
WIDTH (centimeters)	.61

But each predictor also has a high correlation with the other predictors:

Predictor	CYLINDER	POWER	SPEED	WEIGHT	LENGTH
CYLINDER	1				
POWER	.86	1			
SPEED	.69	.89	1		
WEIGHT	.90	.75	.49	1	
LENGTH	.86	.69	.53	.92	1
WIDTH	.71	.55	.36	.79	.86

For COREXpress Tutorial #1 see: statisticalinnovations.com

Example 1: OLS Results

Results from traditional OLS regression: CV-R2 = 0.63

OLS Regression	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
CYLINDER	**-1.9**	33.6	**-.02**	-.06	.95
POWER	1315.9	613.5	.89	2.14	.05
SPEED	**-472.5**	740.3	**-.21**	-.64	.53
WEIGHT	45.9	100.0	.18	.46	.65
LENGTH	209.6	504.2	.15	.42	.68
WIDTH	**-505.4**	1501.6	**-.07**	-.34	.74
(Constant)	12070.4	194786.6		.06	.95

- Since solution is based on a relatively small sample (N=24) and correlated predictors, it is likely that this model overfits the data:

While R^2 = .85, Cross validated R^2 = .63

- OLS solution yields large standard errors and unrealistic negative coefficients for predictors CYLINDER, SPEED, and WIDTH.

CCR Results (K=2 components): CV-R^2= .75

Predictor	B	Beta
CYLINDER	20.9	0.19
POWER	545.5	0.37
SPEED	445.7	0.20
WEIGHT	43.4	0.17
LENGTH	32.6	0.02
WIDTH	343.6	0.05
(Constant)	-177941	

CCR Results **with** Stepdown

CV- R^2 =	0.77	
Predictor	B	Beta
POWER	673.3	0.45
SPEED	222.9	0.10
WEIGHT	110.9	0.44
(Constant)	-115044	

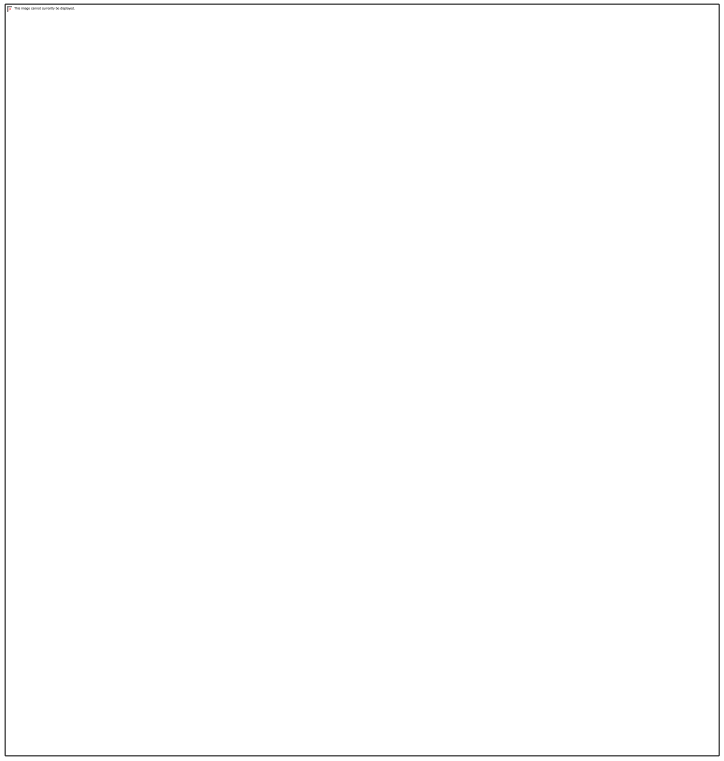
CCR yields more reasonable results: positive coefficients for all 6 predictors.

CCR step-down procedure obtains even better results: (CV-R^2 = .77)

*Based on 10 rounds of 6-fold cross-validation

Example 1: K = 2-component CCR Model

CV-R^2 as a function of # predictors



Training R^2	0.84
CV-R^2	0.77 (.03)

Predictors	Standardized Coefficient
POWER	0.45
WEIGHT	0.44
SPEED	0.10

Predictor	All	1	2	3	4	5	6	7	8	9	10
POWER	60	6	6	6	6	6	6	6	6	6	6
WEIGHT	59	6	6	6	6	6	5	6	6	6	6
SPEED	27	3	6	3	3	2	0	3	4	0	3
CYLINDER	23	2	6	3	2	3	1	3	1	0	2
LENGTH	10	1	6	0	0	1	0	0	1	0	1
WIDTH	7	0	6	0	1	0	0	0	0	0	0
Total	186	18	36	18	18	18	12	18	18	12	18
Predictors		3	6	3	3	3	2	3	3	2	3

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

CCR Variants in COREXpress®

For dependent variables that are not continuous:

CCR-Logistic: Logistic Regression Models

CCR-LDA: Linear Discriminant Analysis

CCR-Cox: Survival (Event History) Models:

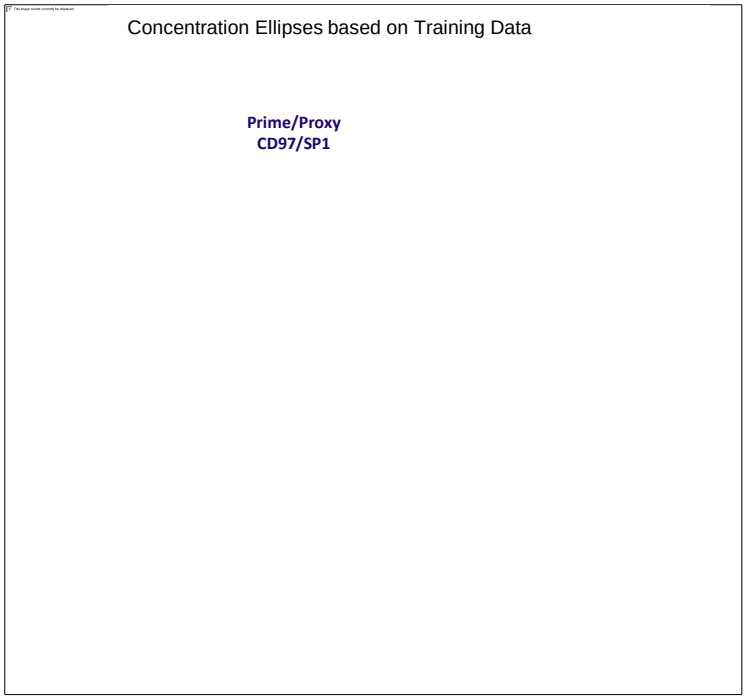
Latent Class (Clustering) Applications:

- Selection of predictors prior to LC modeling
- Separate models for each LC segment

CCR Optimized to Capture Suppressor Effects

Despite the extensive literature documenting the strong enhancement effects of suppressor variables, **most pre-screening methods omit suppressors prior to model development, resulting in suboptimal models.** This is akin to: **"throwing out the baby with the bath water".**

Inclusion of SP1 (suppressor) improves prediction of Cancer (Z=1) vs. Normals (Z=0) over CD97 alone: AUC = .87 vs. .70 (training data), and .84 vs. .73 (validation data). See Magidson and Wassmann (2010).



Cancer Subjects have elevated CD97 level as compared to Normals – Red ellipse lies *above* blue ellipse.

SP1 is a suppressor: Cancer and Normals do not differ on SP1, despite its high correlation with CD97.

Outline of Topics

- Problem: High Dimensional Data in Regression / Multicolinearity
- Solution: Correlated Component Regression (CCR) for Linear Regression
- Sparsity: Step-down Algorithm for Variable Reduction
- Example #1: Predicting Auto Prices
- Extension of CCR to Logistic Regression / Linear Discriminant Analysis (LDA)
  - Importance of Suppressor Variables
- Example #2: LDA Simulated Data Example

Example #2: Linear Discriminant Analysis

Predictors	Unstandardized	Standardized	Importance	Importance Rank
SP1	-9.55	-5.72	5.72	1
GSK3B	4.56	2.48	2.48	2
RB1	-3.82	-2.30	2.30	3
IQGAP1	3.35	2.13	2.13	4
BRCA1	-2.13	-1.36	1.36	5
TNF	2.24	1.32	1.32	6
CDKN1A	2.33	1.29	1.29	7
MAP2K1	2.75	1.20	1.20	8
MYC	-1.81	-1.19	1.19	9
EP300	-1.78	-1.15	1.15	10
CD44	1.85	1.03	1.03	11
CD97	1.44	0.92	0.92	12
SIAH2	1.15	0.87	0.87	13
MAPK1	1.64	0.79	0.79	14
RP5	1.94	0.76	0.76	15
S100A6	1.22	0.74	0.74	16
ABL1	1.44	0.73	0.73	17
NFKB1	1.22	0.70	0.70	18
MTF1	-1.01	-0.62	0.62	19
CDK2	1.20	0.61	0.61	20
IL18	-0.79	-0.56	0.56	21
PTPRC	-0.98	-0.53	0.53	22
SMAD3	-0.57	-0.35	0.35	23
C1QA	-0.29	-0.30	0.30	24
TP53	0.45	0.26	0.26	25
CDKN2A	-0.31	-0.23	0.23	26
CCNE1	-0.21	-0.19	0.19	27
ST14	-0.18	-0.14	0.14	28

- Data simulated according to assumptions of Linear Discriminant Analysis.
- P = G1 + G2 + G3 predictors:
  - G1 = 28 valid predictors (nonzero population coefficients given in Table 1), including 15 relatively weak predictors (valid predictors with importance scores < .85)
  - G2 = 28 irrelevant predictors ('extra1' – 'extra28') uncorrelated with both dependent variable and with the 28 valid predictors but correlated with each other
  - G3 = 28 irrelevant predictors ('INDPT1' – 'INDPT28'), each uncorrelated with all other variables
- Correlations and variances mimic real data.
- 100 simulated samples, each consisting of N=50 cases, with group sizes N1 = N2 = 25.

Statistical Innovations
Thinking outside the brackets!™
logit research

Results from Full CCR-LDA Simulation 100 simulated samples

Method	Misclassification error rate	Number (%) irrelevant variables		% of simulated samples where important suppressor variable included in model	Average # predictors in model
		#	%		
CCR	17.4%	3.4	23.0%	91.0%	14.5
Sparse PLS	19.3%	6.9	34.0%	78.0%	20.4
Elastic Net	21.1%	6.6	34.0%	61.0%	19.2
Lasso	21.6%	4.3	31.0%	51.0%	13.6

Sparse Regression Methods

Correlated Component Regression (CCR), sparse PLS regression (sgpls, Chun and Keles, 2009) , Elastic Net (L1 + L2 regularisation, Zou and Hastie, 2005), and Lasso (L1 regularisation)

Theory: Why CCR Outperforms Traditional Methods

- For data generated according to LDA assumptions, LDA itself does not work well with high dimensional data. In particular, the simple Naïve Bayes Rule:

  > *"greatly outperforms the Fisher linear discriminant rule (LDA) under broad conditions when the number of variables grows faster than the number of observations"*, Bickel and Levina (2004)

- Naïve Bayes rule is equivalent to the 1-component CCR model (CCR1).
- Traditional regression is equivalent to a saturated CCR model – CCR with at most K=min(P,N-1) components.
- Typically, CCR with 2-8 components (CCR2-CCR8) works best in practice.

Note: Naïve Bayes fails to capture the effects of suppressor variables since by definition, suppressor variables will have zero loadings on component #1.

References

- Magidson, Jay (2012, forthcoming). "Correlated Component Regression: Re-thinking Regression in the Presence of Near Collinearity", in Springer Verlag series: "New perspectives in Partial Least Squares and Related Methods".
- Magidson and Wassmann (2010). "The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer", Proceedings from the 2010 Joint Statistical Meetings.
- Statistical Innovations (2011). COREXpress® User's Guide.

